Modeling missing-data processes: An IRTree-based approach.

Dries Debeer (University of Leuven)
Rianne Janssen (University of Leuven)
Paul De Boeck (Ohio State University)

This talk?

Modeling missing item responses ..

.. in low-stake assessments ..
(educational measurement)

.. using IRTrees.
(De Boeck & Partchev, 2012)

Content

1. Missing data
2. Handling missing data
3. IRTree
4. IRTrees for missing data
5. Illustration
6. Discussion

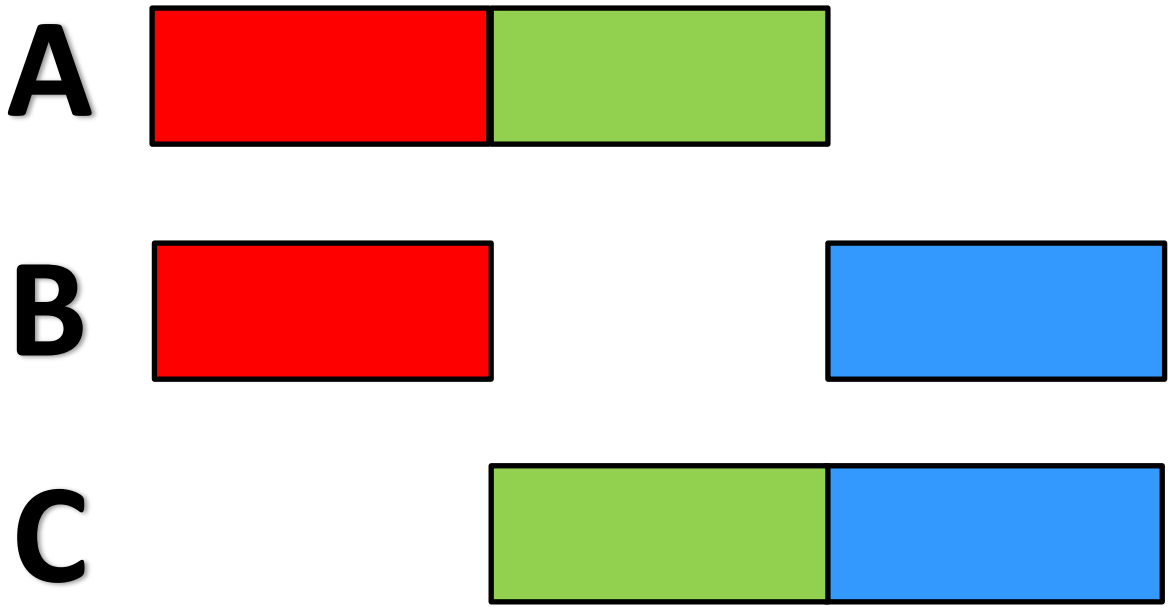
Missing data

Different types

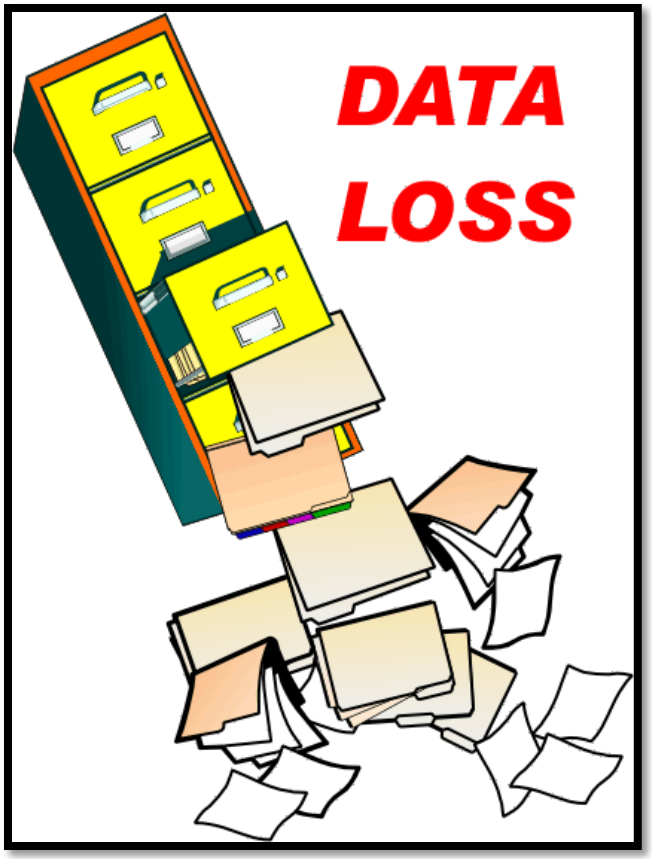
1. Before administration
2. After administration
3. During administration

1. Before administration

By Design



2. After administration



Missing data

Different types

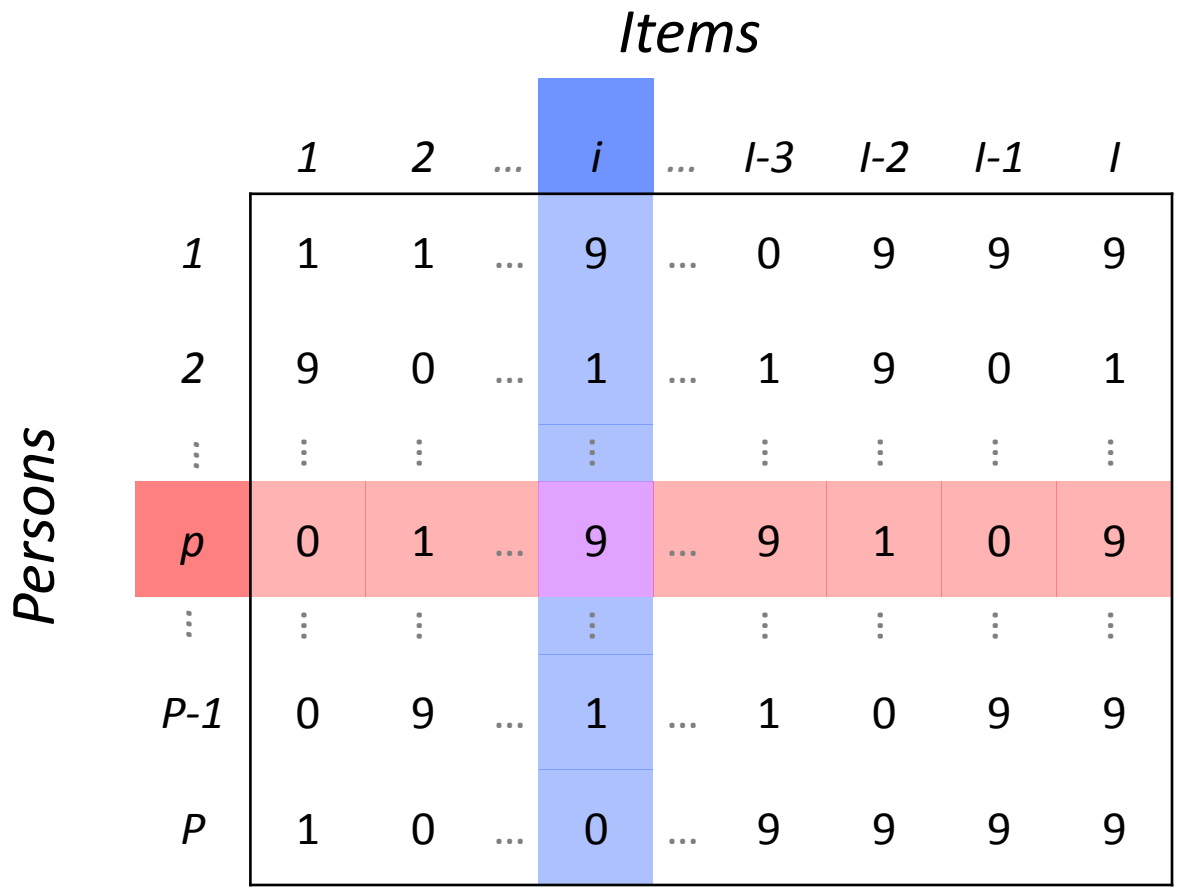
1. Before administration
2. After administration
3. During administration
   - Related to test takers
   - Related to items

3. During administration

Two types of non-responses:

1. Near the end of the assessment
   - “r”
   - “not reached”
2. Well before the end of the assessment
   - “d”
   - “skipped”

Data Matrix: $\boldsymbol{X}$ $(P \times I)$



Data Matrix: X

	1	*2*	…	*i*	…	*I-3*	*I-2*	*I-1*	*I*
1	1	1	…	**9**	…	0	**9**	**9**	**9**
2	**9**	0	…	1	…	1	**9**	0	1
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
p	0	1	…	**9**	…	**9**	1	0	**9**
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
P-1	0	**9**	…	1	…	1	0	**9**	**9**
P	1	0	…	0	…	**9**	**9**	**9**	**9**

Missing responses: “9”

Data Matrix: X

	1	*2*	…	*i*	…	*I-3*	*I-2*	*I-1*	*I*
1	1	1	…	**d**	…	0	**9**	**9**	**9**
2	**d**	0	…	1	…	1	**d**	0	1
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
p	0	1	…	**d**	…	**d**	1	0	**9**
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
P-1	0	**d**	…	1	…	1	0	**9**	**9**
P	1	0	…	0	…	**9**	**9**	**9**	**9**

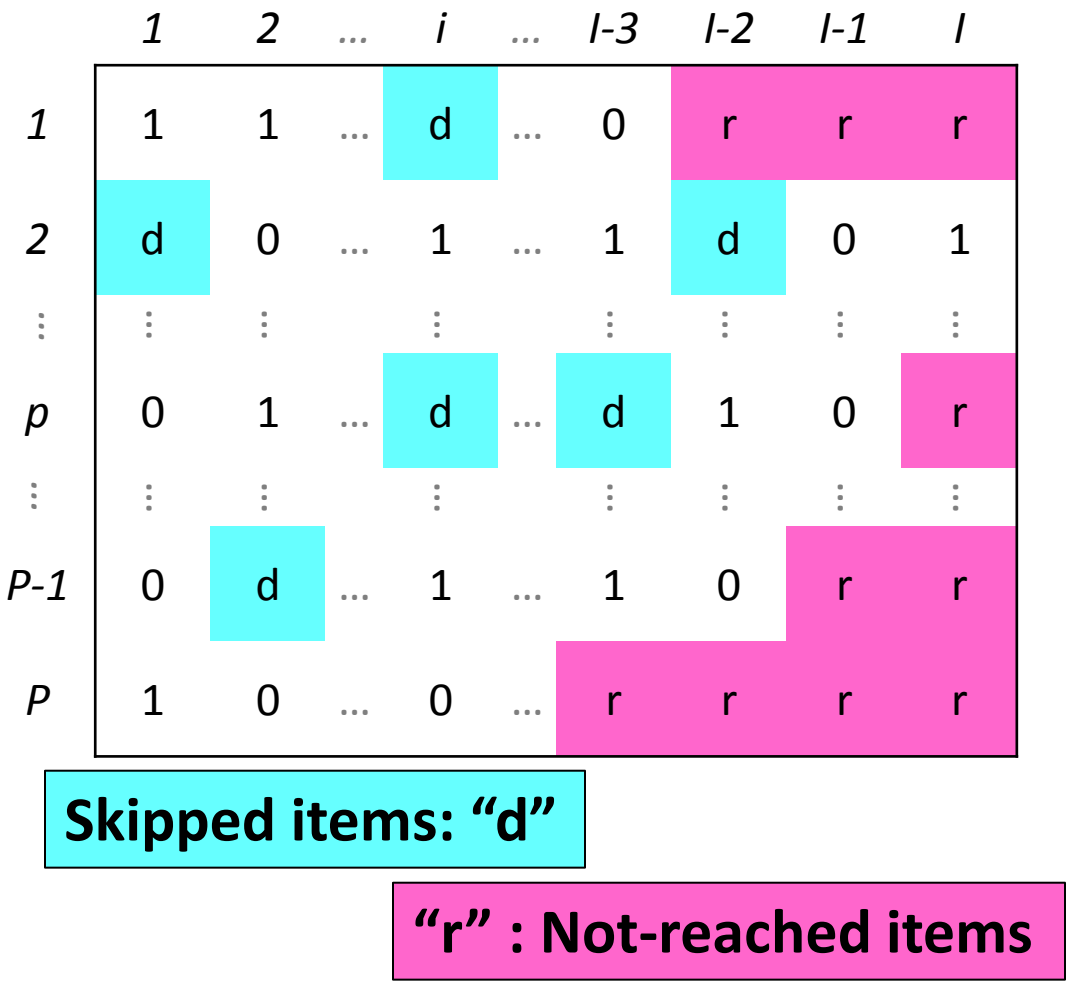
Skipped items: “d”

Data Matrix: X

	1	*2*	…	*i*	…	*I-3*	*I-2*	*I-1*	*I*
1	1	1	…	d	…	0	**r**	**r**	**r**
2	d	0	…	1	…	1	d	0	1
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
p	0	1	…	d	…	d	1	0	**r**
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
P-1	0	d	…	1	…	1	0	**r**	**r**
P	1	0	…	0	…	**r**	**r**	**r**	**r**

Not-reached items: “r”

Data Matrix: X



Content

1. Missing data
2. Handling missing data
3. IRTree
4. IRTrees for missing data
5. Illustration
6. Discussion

Handling missing data

Common strategies:

- Missing = Wrong
  - ***r*/*d* = *0***
  - Data imputation
- Ignore missing responses
  - ***r*/*d* = *NA***
  - Assuming MAR
- Combination of both

Handling missing data

However…

Missingness might be related to

- Lower *motivation* => higher probability of ***d***?
- Lower *speed* => higher probability of ***r***?
- …

$\Rightarrow$ The ***r***/***d*** -responses can be informative.

Missingness might be related to

- the proficiency of the test taker.

$\Rightarrow$ Bias on the ability measurement (MNAR)

$\Rightarrow$ **Not ignorable**

Missing data: Categorisation (Rubin, 1976)

MCAR = Missing Completely At Random

MAR = Missing At Random

MNAR = Missing Not At Random

Missingness indicator u_{pi}

- Whenever x_{pi} is observed, $u_{pi} = 1$
- Whenever x_{pi} is not observed, $u_{pi} = 0$
- $\boldsymbol{U}$: $P \times I$ $matrix$

$\boldsymbol{X}_{obs}$ & $\boldsymbol{X}_{mis}$

Based on the missingness indicator matrix $\boldsymbol{U}$ the *full* data matrix falls apart in:

- observed part $\boldsymbol{X}_{obs}$
- unobserved part $\boldsymbol{X}_{mis}$

MCAR

The missingness does not depend on the observed nor the unobserved data.

$$P(U|X_{obs}, X_{mis}) = P(U)$$

- No bias
- No problems for analyses

MAR

The missingness does only depend on the observed data.

$$P(U|X_{obs}, X_{mis}) = P(U|X_{obs})$$

- No bias for likelihood-based inferences
- When MCAR, also MAR

MNAR

The missingness does depend on the observed and the unobserved data.

$$P(U|X_{obs}, X_{mis}) = P(U|X_{obs}, X_{mis})$$

- **Non–ignorable** for likelihood based inferences
- Possible bias

Model missing data process/processes

Modeling missing data processes

Latent variable model (Moustaki & O'Muircheartaigh, 2000)

- Skipped (Holman & Glas, 2005)
- Not reached (Glas & Pimentel, 2008)
- Both…

$\Rightarrow$ IRTree (De Boeck & Partchev, 2012)

Content

1. Missing data
2. Handling missing data
3. IRTree
4. IRTrees for missing data
5. Illustration
6. Discussion

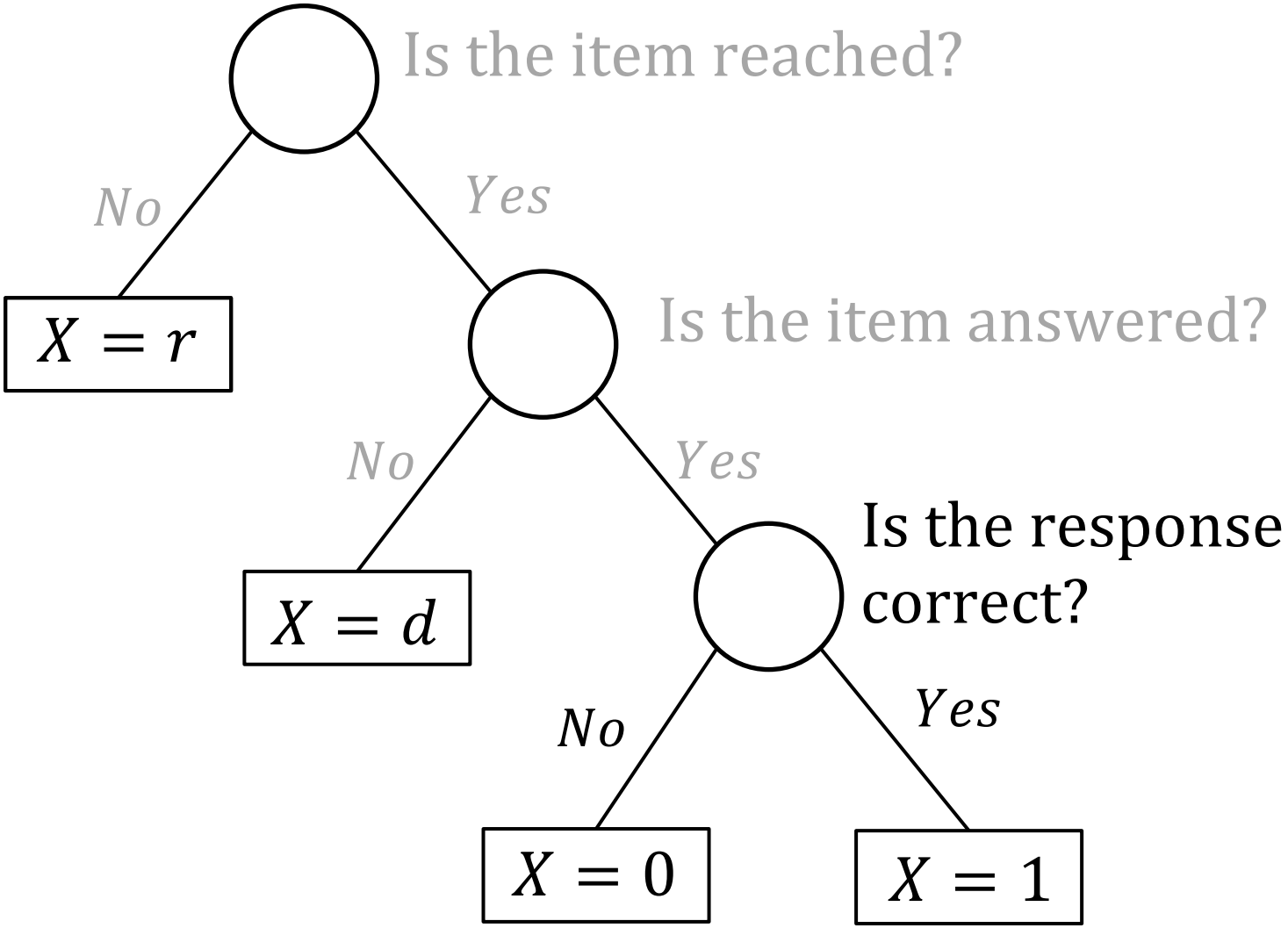
IRTree (De Boeck & Partchev, 2012)

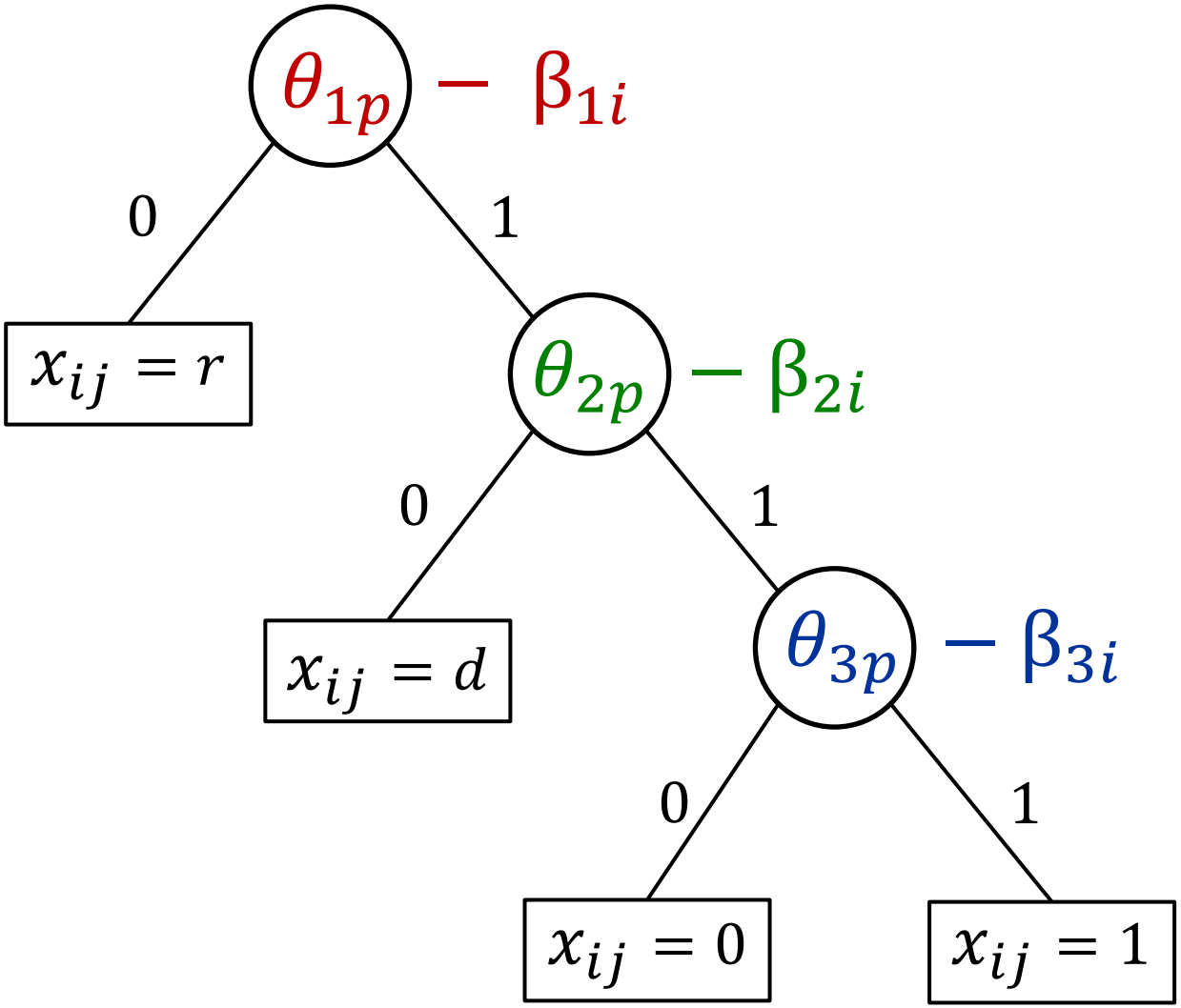
- IRT-model
- tree structure
- sequentially interconnected sub-processes
- model a response process
  - observed variables (responses) 
  - unobserved (latent) sub-processes 

Here:

- IRTree to model response categories $(0, 1, d, r)$
- With distinct missing data processes
- 2 IRTrees

IRTree 1


Is the item reached?
No
Yes
$X = r$
Is the item answered?
No
Yes
$X = d$
Is the response correct?
No
Yes
$X = 0$
$X = 1$


$\theta_{1p} - \beta_{1i}$
0
1
$x_{ij} = r$
$\theta_{2p} - \beta_{2i}$
0
1
$x_{ij} = d$
$\theta_{3p} - \beta_{3i}$
0
1
$x_{ij} = 0$
$x_{ij} = 1$

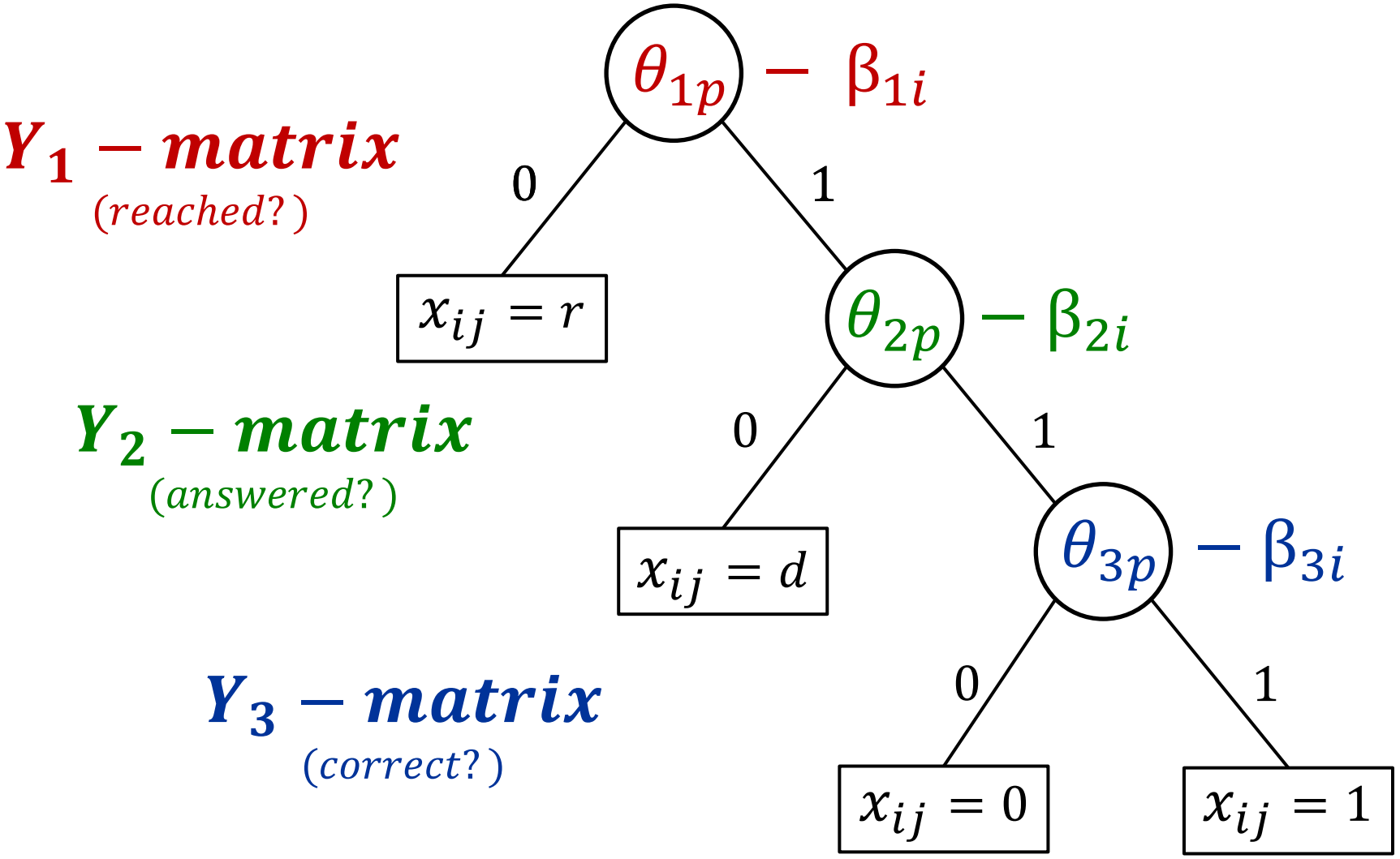
IRTree 1: Probabilities

$$P[x_{ij} = r] = 1 - logit^{-1}[\theta_{1p} - \beta_{1i}]$$

$$P[x_{ij} = d] = logit^{-1}[\theta_{1p} - \beta_{1i}] \times (1 - logit^{-1}[\theta_{2p} - \beta_{2i}])$$

$$P[x_{ij} = 0] = logit^{-1}[\theta_{1p} - \beta_{1i}] \times logit^{-1}[\theta_{2p} - \beta_{2i}] \times (1 - logit^{-1}[\theta_{3p} - \beta_{3i}])$$

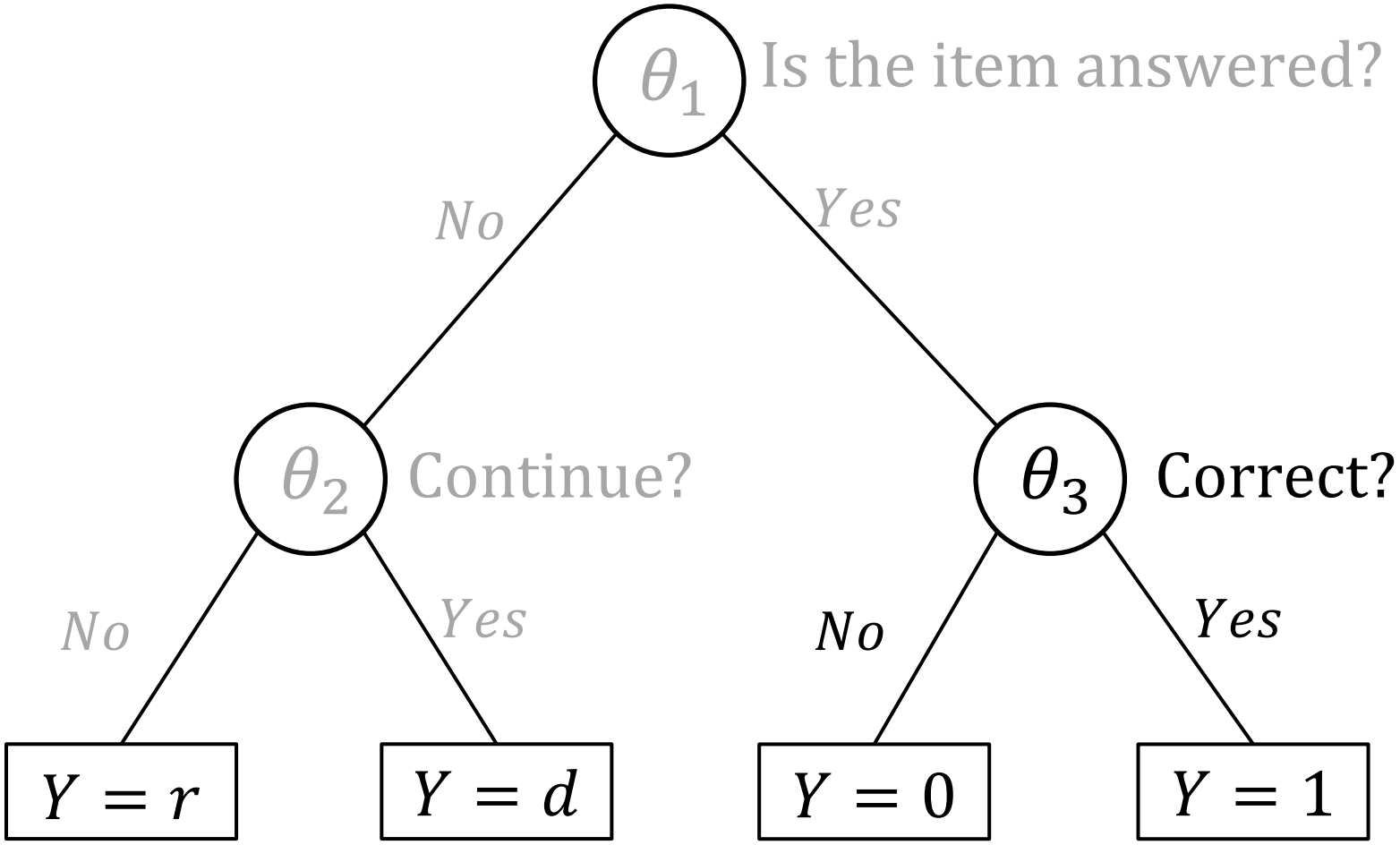
$$P[x_{ij} = 1] = logit^{-1}[\theta_{1p} - \beta_{1i}] \times logit^{-1}[\theta_{2p} - \beta_{2i}] \times logit^{-1}[\theta_{3p} - \beta_{3i}]$$


$Y_1 - matrix$
(reached?)
$Y_2 - matrix$
(answered?)
$Y_3 - matrix$
(correct?)
$\theta_{1p} - \beta_{1i}$
0
1
$x_{ij} = r$
$\theta_{2p} - \beta_{2i}$
0
1
$x_{ij} = d$
$\theta_{3p} - \beta_{3i}$
0
1
$x_{ij} = 0$
$x_{ij} = 1$

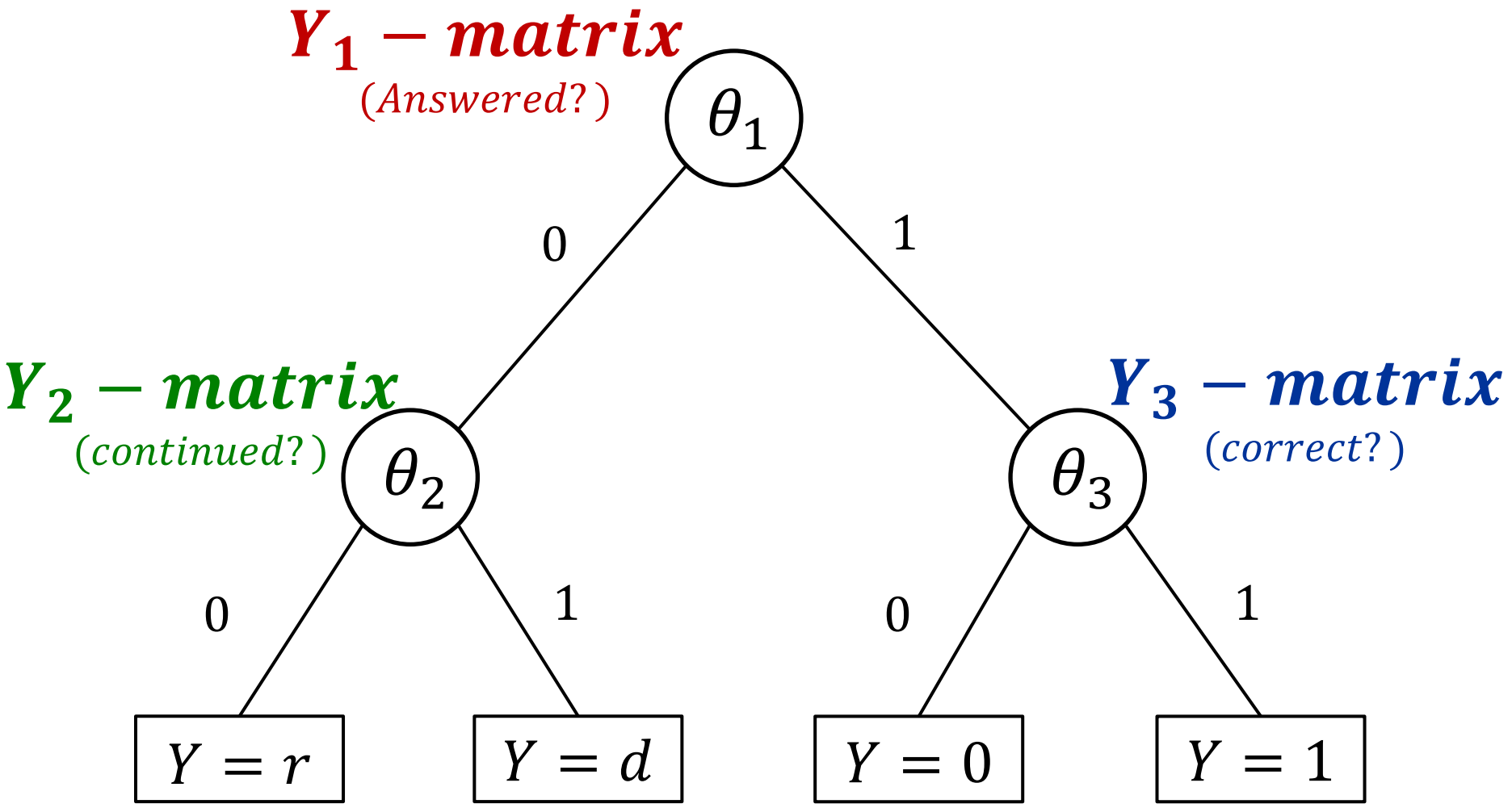
Mapping on subitems

Item	X_{obs}	*Code*	Y_1	Y_2	Y_3
1	1	1	**1**	**1**	**1**
2	1	0	**1**	**1**	**1**
3	9	d	**1**	**0**	**NA**
4	0	0	**1**	**1**	**0**
5	9	r	**0**	**NA**	**NA**
6	9	r	**NA**	**NA**	**NA**
7	9	r	**NA**	**NA**	**NA**

IRTree 2


θ_1
Is the item answered?
No
Yes
θ_2
Continue?
θ_3
Correct?
No
Yes
No
Yes
$Y = r$
$Y = d$
$Y = 0$
$Y = 1$

IRTree 2: Subitems



IRTree 2: Probabilities

$$P[x_{ij} = r] = \left(1 - logit^{-1}[\theta_{1p} - \beta_{1i}]\right) \times \left(1 - logit^{-1}[\theta_{2p} - \beta_{2i}]\right)$$

$$P[x_{ij} = d] = \left(1 - logit^{-1}[\theta_{1p} - \beta_{1i}]\right) \times logit^{-1}[\theta_{2p} - \beta_{2i}]$$

$$P[x_{ij} = 0] = logit^{-1}[\theta_{1p} - \beta_{1i}] \times \left(1 - logit^{-1}[\theta_{3p} - \beta_{3i}]\right)$$

$$P[x_{ij} = 1] = logit^{-1}[\theta_{1p} - \beta_{1i}] \times logit^{-1}[\theta_{3p} - \beta_{3i}]$$

Mapping on subitems

Item	X_{obs}	*Code*	Y_1	Y_2	Y_3
1	1	1	**1**	**NA**	**1**
2	1	0	**1**	**NA**	**1**
3	9	d	**0**	**1**	**NA**
4	0	0	**1**	**NA**	**0**
5	9	r	**0**	**0**	**NA**
6	9	r	**NA**	**NA**	**NA**
7	9	r	**NA**	**NA**	**NA**

1PL IRTree

$$\pi\left(x_{pi} = m \middle| \boldsymbol{\theta}_p, \boldsymbol{\beta}_i\right) = \prod_{r=1}^{R} \left[\frac{\exp(\theta_p^{(n)} + \beta_i^{(n)})^{t_{mn}}}{1 + \exp(\theta_p^{(n)} + \beta_i^{(n)} \beta_i^{(n)})} \right]^{d_{mn}}$$

$m = (0, 1, d, r)$ (response categories)

$n = (1, 2, 3)$ (nodes)

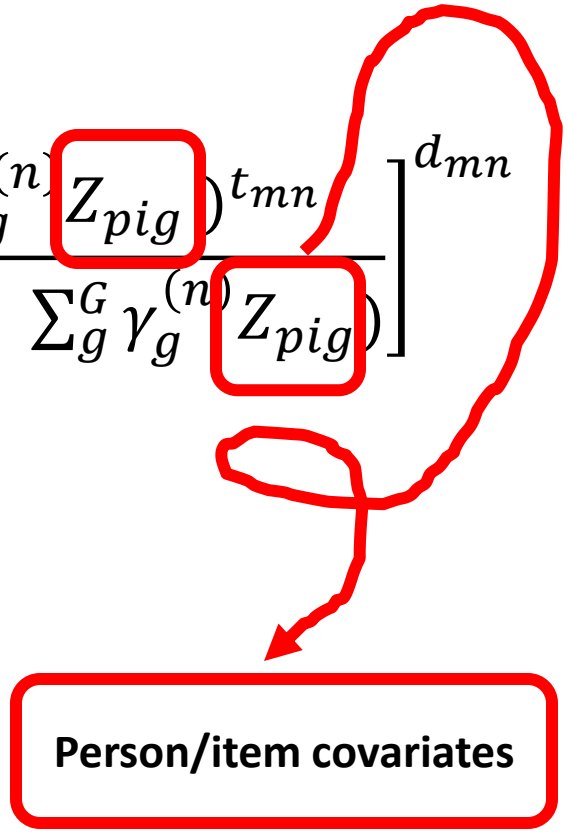
$t_{mn} = (0,1, NA)$ mapping of m on n

$$d_{mn} = \begin{cases} 0 \text{ when } t_{mr} = NA \\ 1 \text{ otherwise} \end{cases}$$

Family of GLMM

IRTrees

$$\pi\left(x_{pi} = m \middle| \boldsymbol{\theta}_p, \boldsymbol{\beta}_i\right) = \prod_{r=1}^{R} \left[\frac{\exp(\theta_p^{(n)} + \beta_i^{(n)} + \sum_g^G \gamma_g^{(n)} Z_{pig})^{t_{mn}}}{1 + \exp(\theta_p^{(n)} + \beta_i^{(n)} \beta_i^{(n)} + \sum_g^G \gamma_g^{(n)} Z_{pig})} \right]^{d_{mn}}$$



$m = (0, 1, d, r)$ (response categories)

$n = (1, 2, 3)$ (nodes)

$t_{mr} = (0, 1, NA)$ mapping of m on n

$$d_{mr} = \begin{cases} 0 \text{ when } t_{mr} = NA \\ 1 \text{ otherwise} \end{cases}$$

IRTrees

Relation between missing data processes can be tested

- Correlation between latent traits
- Correlation between item parameters
- MNAR
- Missing = wrong
- Multinomial model

Interpretation IRTree

Proficiency process: the same

Missing data processes: different

IRTree1: *Item selection model* ?

- For more speeded tests?

IRTree2: *Continuing effort model* ?

- Motivation / fatigue?

Which IRTree?

- Proficiency
- Test
- Administering conditions

- Model fit (information criterion)

Content

1. Missing data
2. Handling missing data
3. IRTree
4. IRTrees for missing data
5. Illustration
6. Discussion

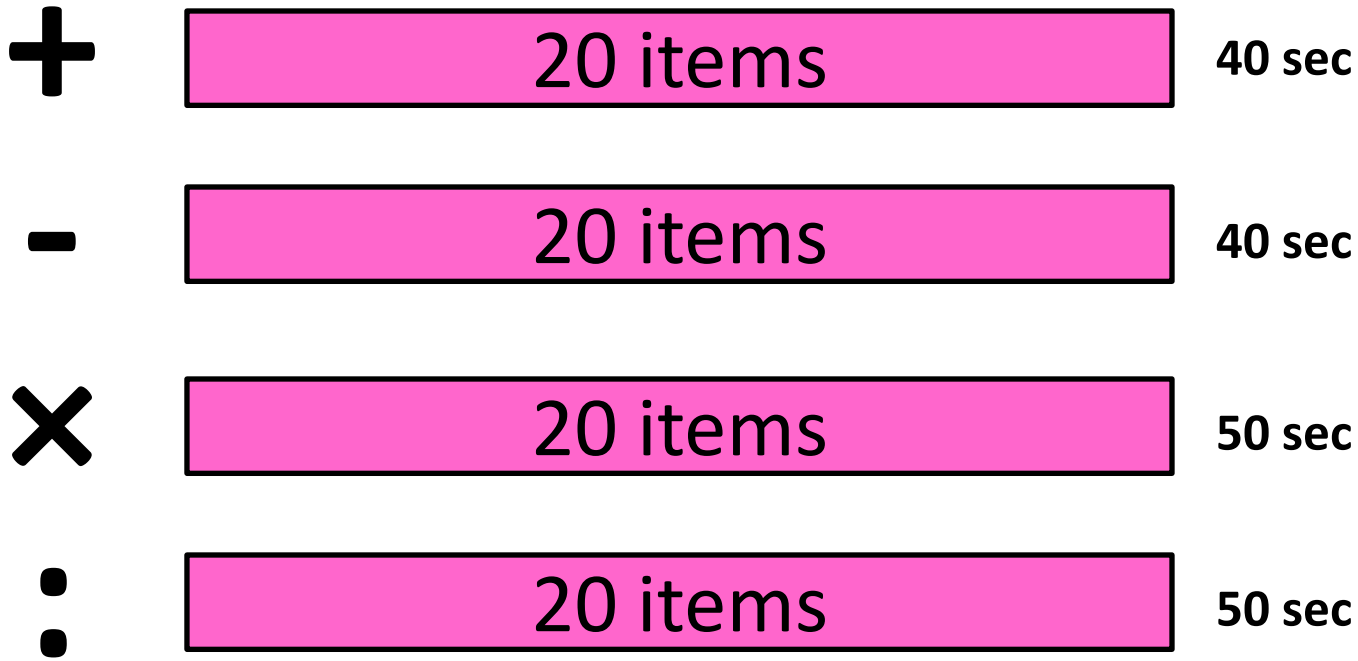
Illustration: Fast arithmetic

Part of national math survey in the Flemisch part of Belgium

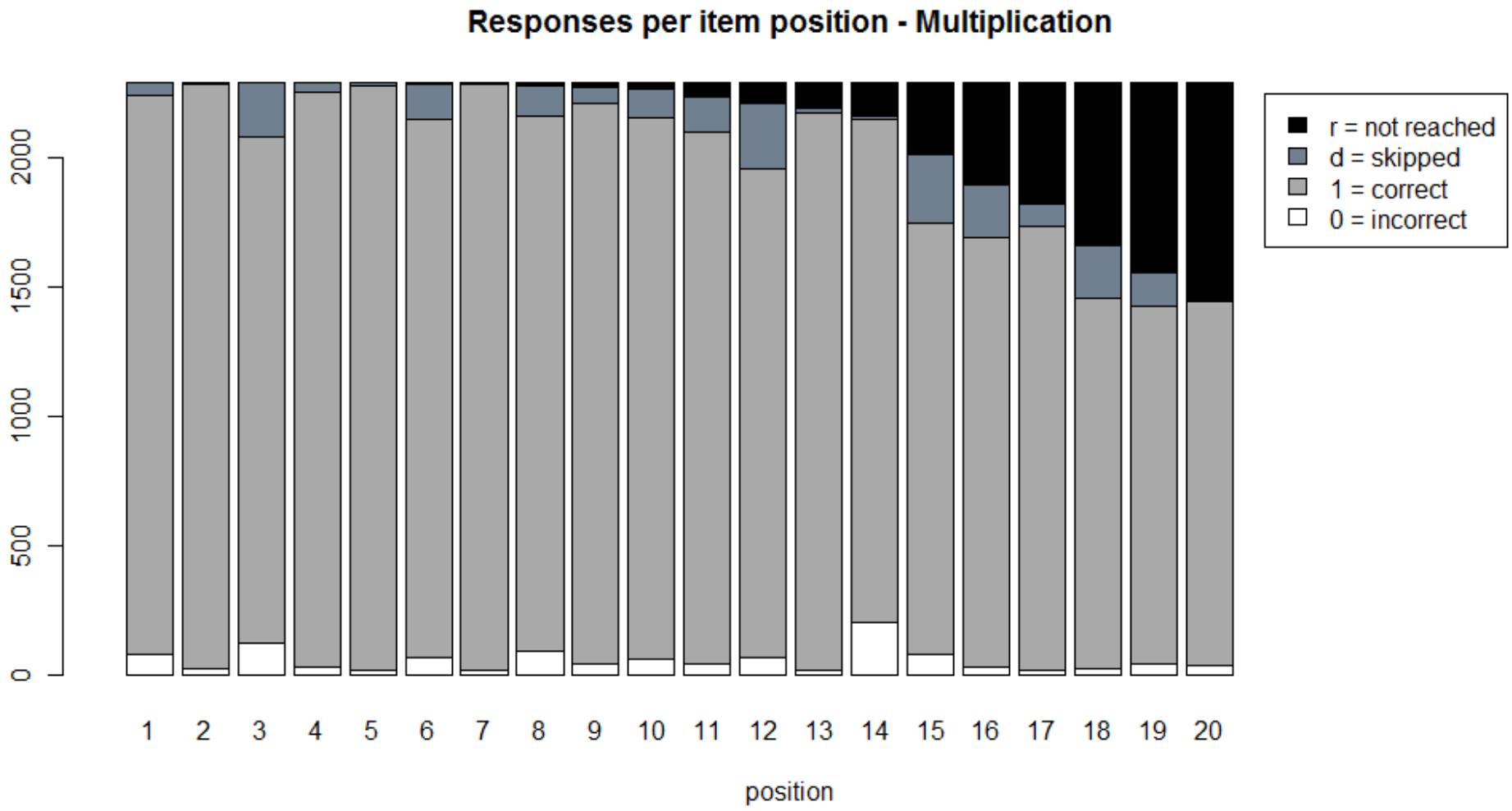
N = 2288 students

Speed test (easy items)

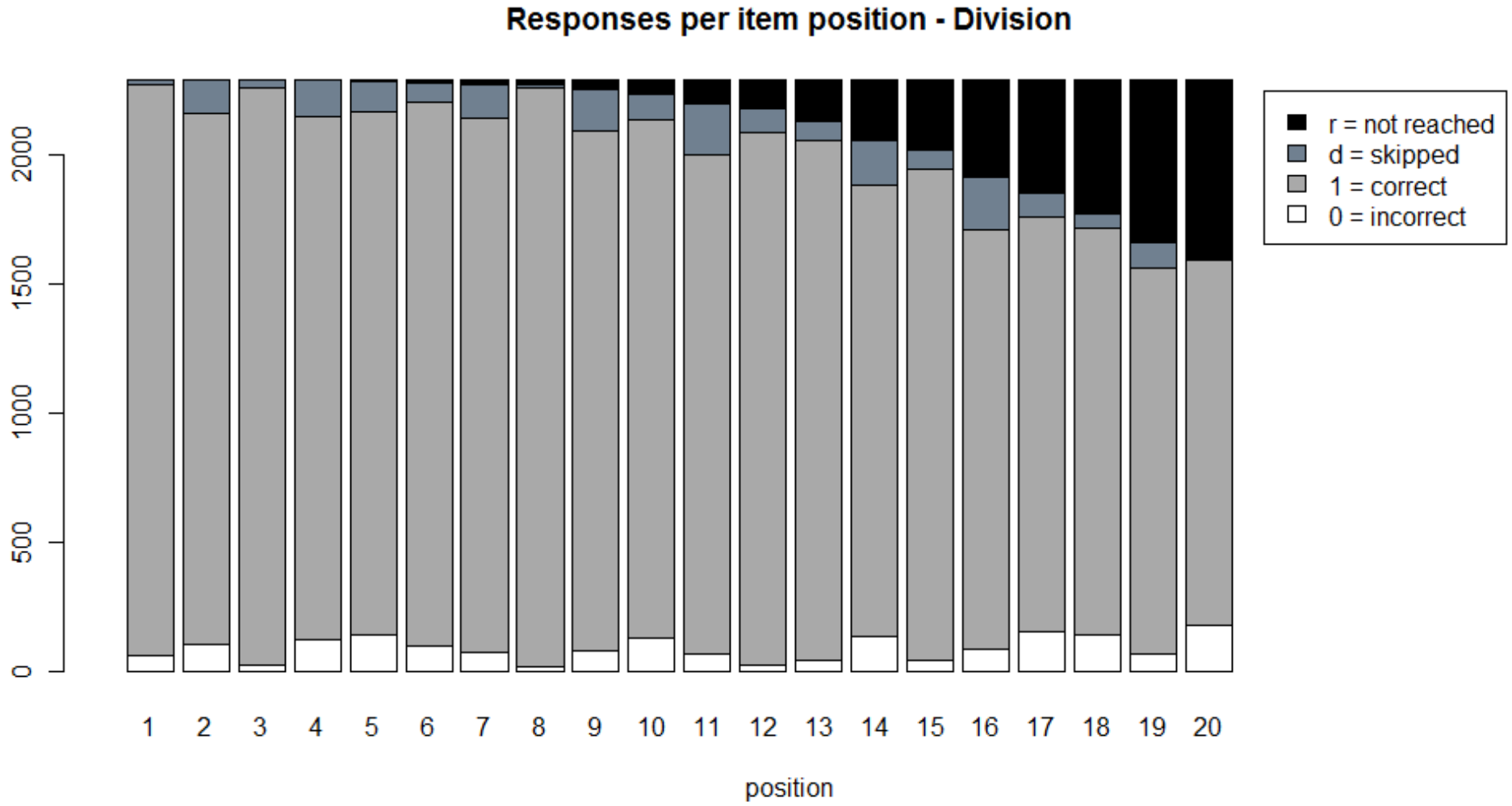
Design



Descriptives



Descriptives



Estimation

In R (Lmer)

$(\theta_p^1, \theta_p^2, \theta_p^3) \sim MNV(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$

$(\varepsilon_i^1, \varepsilon_i^2, \varepsilon_i^3) \sim MNV(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$

$\Rightarrow$ person / item correlations are estimated

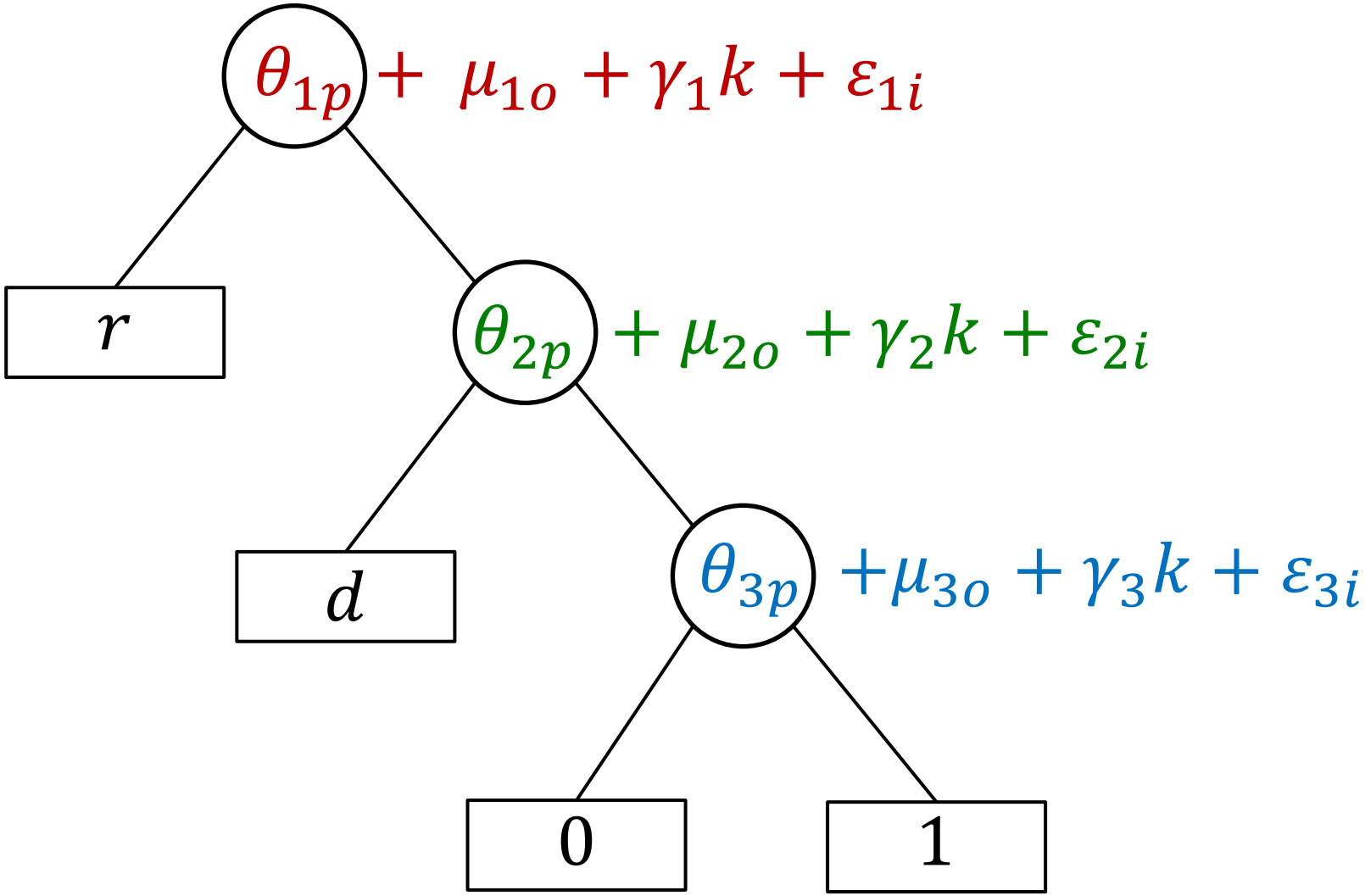
$\Rightarrow$ relation between the processes?

Which IRTree?

Speed tests: IRTree1 (*Item selection model)*

Model	- logL	AIC	BIC
IRTree1	44137	88314	88516
IRTree2	44563	89167	89369

Final model


$\theta_{1p} + \mu_{1o} + \gamma_1 k + \varepsilon_{1i}$
r
$\theta_{2p} + \mu_{2o} + \gamma_2 k + \varepsilon_{2i}$
d
$\theta_{3p} + \mu_{3o} + \gamma_3 k + \varepsilon_{3i}$
0
1

Results

	Position effect		
	γ_n	SE(γ_n)	p
Node 1	-0.461	.0143	< .0001
Node 2	-0.055	. 0245	.0256
Node 3	-0.038	.0148	.0099

Results

Correlations:

$$\begin{matrix}\widehat{\theta_1} \\ \widehat{\theta_2} \\ \widehat{\theta_3}\end{matrix}\begin{pmatrix}1.98 & & \\ 0.57 & 2.41 & \\ 0.53 & 0.56 & 0.86\end{pmatrix}$$

$$\begin{matrix}\widehat{\varepsilon_1} \\ \widehat{\varepsilon_2} \\ \widehat{\varepsilon_3}\end{matrix}\begin{pmatrix}0.42 & & \\ 0.69 & 1.21 & \\ 0.29 & 0.31 & 0.74\end{pmatrix}$$

Results

- Clear effect of item position on *reaching*-threshold (design with different item orders is to investigate item position effects on difficulty)
- Missing data processes are related, but not the same as proficiency process.
  - MNAR
  - $\boldsymbol{r/d \neq 0}$
- Threshold for reaching and skipping are related.

Further..

- Polytomous items / 2PL
- Apply to PISA data
  - Differences between countries?
  - Impact on country ranking?

Discussion

- Information in missing responses
- A lot of assumptions
- How much missing data is needed to get reliable estimates?
- Applicable in other domains?
- How to interpret the missingness processes?
  - Related to motivation,.. ?
  - Related to response times?
- When desirable?
  - Interested in missing data processes
  - Using information in missing responses

To remember

- Missing responses can be non-ignorable.
- But they can be modeled.
- We use IRTrees
  - $\Rightarrow$ Can reduce possible bias
  - $\Rightarrow$ Can help in understanding the processes underlying the missing data.

Thank you for your attention.